

Lenient Multi-Agent Deep Reinforcement Learning

Gregory Palmer

Department of Computer Science
University of Liverpool, United Kingdom
G.J.Palmer@liverpool.ac.uk

Karl Tuyls

DeepMind and University of Liverpool
United Kingdom
karltuyls@google.com

Daan Bloembergen

Department of Computer Science
University of Liverpool, United Kingdom
D.Bloembergen@liverpool.ac.uk

Rahul Savani

Department of Computer Science
University of Liverpool, United Kingdom
Rahul.Savani@liverpool.ac.uk

Abstract:

A significant amount of research in recent years has been dedicated towards single agent deep reinforcement learning. Much of the success of deep reinforcement learning can be attributed towards the use of experience replay memories within which state transitions are stored. Function approximation methods such as convolutional neural networks (referred to as deep Q-Networks, or DQNs, in this context) can subsequently be trained through sampling the stored transitions. However, considerations are required when using experience replay memories within multi-agent systems, as stored transitions can become outdated due to agents updating their respective policies in parallel [1]. In this work we apply *leniency* [2] to multi-agent deep reinforcement learning (MA-DRL), acting as a control mechanism to determine which state-transitions sampled are allowed to update the DQN. Our resulting Lenient-DQN (LDQN) is evaluated using variations of the Coordinated Multi-Agent Object Transportation Problem (CMOTP) outlined by Buşoniu et al. [3]. The LDQN significantly outperforms the existing hysteretic DQN (HDQN) [4] within environments that yield stochastic rewards. Based on results from experiments conducted using vanilla and double Q-learning versions of the lenient and hysteretic algorithms, we advocate a hybrid approach where learners initially use vanilla Q-learning before transitioning to double Q-learners upon converging on a cooperative joint policy.

1 Introduction

The field of *deep reinforcement learning* has seen a great number of successes in recent years. Deep reinforcement learning agents have been shown to master numerous complex problem domains, ranging from computer games [5, 6, 7, 8] to robotics tasks [9, 10]. Much of this success can be attributed to using convolutional neural network (*ConvNet*) architectures as function approximators, allowing algorithms from traditional reinforcement learning to be applied to domains that suffer from the curse of dimensionality. ConvNets are often trained to approximate policy and value functions through sampling past state transitions stored by the agent inside an experience replay memory.

Recently the sub-field of multi-agent deep reinforcement learning (MA-DRL) has received an increased amount of attention. One of the key challenges faced within multi-agent reinforcement learning is the *moving target problem*: Given an environment with multiple agents whose rewards depend on each others' actions, then the difficulty of finding optimal policies for each agent is increased due to the policies of the agents being non stationary [11, 12]. The use of an experience

replay memory amplifies this problem, as a large proportion of the state transitions stored inside an agent’s experience replay memory can become deprecated.

A number of methods have been proposed to help deep reinforcement learning agents converge towards an optimal joint policy inside a multi-agent environment. Gupta et al. [13] evaluated policy gradient, temporal difference error and actor critic methods on cooperative control tasks, that included discrete and continuous state and action spaces, using a decentralized parameter sharing approach with centralized learning. In contrast our current work focuses on concurrent learning. A recent successful approach has been to decompose a team value function into agent-wise value functions through the use of a value decomposition network architecture [14]. Others have attempted to help concurrent learners converge through identifying and deleting obsolete state transitions stored inside the replay memory. Foerster et al. [1] for instance used importance sampling as a means to identify outdated transitions while maintaining an action observation history of the other agents. Our current work does not require the agents to maintain an action observation history, nor do we attempt to delete obsolete transitions. Instead we focus on optimistic agents.

Recently Omidshafiei et al. [4] successfully applied concepts from hysteretic Q-learning to MA-DRL. Hysteretic Q-learning is a form of optimistic learning that uses two different learning rates: a higher learning rate for updates that increase the estimate of the value of a state-action pair (Q-value) and a smaller learning rate for updates that decrease the Q-value. The success of the hysteretic approach raises the question whether the leniency concept [2] can be applied to MA-DRL. Both leniency and hysteretic Q-learning are well researched approaches from traditional reinforcement learning that are employed to help parallel learning agents converge towards an optimal joint policy in cooperative settings. Similar to the hysteretic approach lenient agents initially adopt an optimistic disposition, before gradually transforming into average reward learners [15].

2 Background

Q-Learning. The algorithms implemented for this study are based upon Q-learning, a form of temporal difference reinforcement learning that is well suited for solving sequential decision making problems that yield stochastic and delayed rewards [16, 17]. The algorithm strives to find quality values (Q-Values) for state-action pairs [16]. Each Q-Value is an estimate of the discounted sum of future rewards that can be obtained at time t through selecting action a_t in a state s_t , providing the optimal policy is selected in each state that follows [17]. Since most interesting sequential decision problems have a large state-action space, Q-values are often approximated using a function approximator such as a neural network or tile coding. The parameters θ of the function approximator can be trained through the learning agent exploring their environment, choosing an action a_t in state s_t according to a policy π , and performing an update based upon the immediate reward r_{t+1} received in state s_{t+1} [7]:

$$\theta_{t+1} = \theta_t + \alpha(Y_t^Q - Q(s_t, a_t; \theta_t))\nabla_{\theta_t} Q(s_t, a_t; \theta_t). \quad (1)$$

Upon performing an update using Equation 1 the Q-Value $Q(s_t, a_t; \theta_t)$ shifts towards the target value Y_t^Q . The α within the equation is a scalar used to control the learning rate, with $\alpha \in (0, 1]$. The target value Y_t^Q is based on the reward received and the highest estimated Q-Value for an action in state s_{t+1} as outlined in Equation 2. A discount rate $\gamma \in (0, 1]$ is applied to this estimate.

$$Y_t^Q \equiv r_{t+1} + \gamma \max_{a \in A} Q(s_{t+1}, a; \theta_t). \quad (2)$$

Deep Q-Networks (DQN). In deep reinforcement learning a multi-layer neural network is used as a function approximator, mapping a set of n-dimensional state variables to a set of m-dimensional Q-Values $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, where m represents the number of actions available to the agent [7]. The network parameters θ can be trained using stochastic gradient descent, randomly sampling past transitions experienced by the agent that are stored within an experience replay memory [5].

Transitions are a tuple consisting of the original state s_t , the action a_t , the resulting state s_{t+1} and the reward r_{t+1} . The network is trained to minimize the time dependent loss function $L_i(\theta_i)$,

$$L_i(\theta_i) = \mathbf{E}_{s,a \sim p(\cdot)} \left[(Y_t - Q(s, a; \theta_i))^2 \right], \quad (3)$$

where $p(s, a)$ represents a probability distribution of the transitions stored within the experience replay memory, and Y_t is the target:

$$Y_t \equiv r_{t+1} + \gamma Q(s_{t+1}, \max_{a \in A} Q(s_{t+1}, a; \theta'_t); \theta'_t). \quad (4)$$

Equation 4 is a form of double Q-learning [18]. The target action is selected using weights θ , while the target value is computed using weights θ' from a target network. The target network is a more stable version of the current network, with the weights being copied from current to target network after every n transitions [8]. Double-DQNs have been shown to reduce overoptimistic value estimates [7]. This notion is interesting for our current work, since both leniency and hysteretic Q-learning attempt to induce sufficient optimism in the early learning phases to allow the learning agents to converge towards an optimal joint policy. As a result our experiments featured an extensive comparison of vanilla and double Q-learning versions of the algorithms studied.

Hysteretic Q-Learning. Hysteretic Q-learning is a robust algorithm that has recently been applied to MA-DRL [4]. Two learning rates are used, α and β , with $\beta < \alpha$. The smaller learning rate β is used whenever an update would reduce a Q-value [19]. Given a spectrum with traditional Q-learning at one end and maximum-based learning at the other, where negative experiences are completely ignored, then hysteretic Q-learning lies somewhere in between depending on β .

Leniency. Lenient learning was originally introduced by Potter and De Jong [20] to help cooperative co-evolutionary algorithms converge towards an optimal policy. It was designed to prevent *relative overgeneralization*, which occurs when agents gravitate towards a sub-optimal joint policy due to each agent's respective policy being the optimal choice when combined with the arbitrary policies selected by the other agents [15]. The same problem arises in repeated games with cooperative independent reinforcement learning agents. However, leniency has been shown to increase the likelihood of convergence towards the globally optimal solution in stateless cooperation games for reinforcement learning agents [2, 21, 16] which should not come as a surprise given that both reinforcement learning and evolutionary algorithms converge towards replicator dynamics of evolutionary game theory [22, 16, 23].

Lenient learners are more likely to converge towards an optimal policy due to forgiving actions by teammates that lead to low rewards during the initial exploration phase [2, 21, 16]. While initially adopting an optimistic disposition the amount of leniency displayed is gradually decayed each time a state-action pair is visited. As a result the agents become average reward learners for frequently visited state-action pairs [15] while remaining optimistic within unexplored areas. The transition to average reward learners helps lenient agents avoid sub-optimal joint policies in environments that yield stochastic rewards [15].

During training the frequency with which lenient reinforcement learning agents perform updates that result in lowering the Q-value of a state action pair (s, a) is determined by leniency and temperature functions, $l(s_t, a_t)$ and $T_t(s_t, a_t)$ respectively. The relation of the temperature function is one to one, with each state-action pair being assigned a temperature value which is decayed each time the pair is visited. The leniency function $l(s_t, a_t) = 1 - e^{-K * T_t(s_t, a_t)}$ uses a constant K as a leniency moderation factor to determine how the temperature value affects the drop-off in leniency [15]. Following the update $T_t(s_t, a_t)$ is decayed using a discount factor $\beta \in [0, 1]$: $T_{t+1}(s_t, a_t) =$

$\beta T_t(s_t, a_t)$. Given a TD-Error δ , where $\delta = Y_t - Q(s_t, a_t; \theta_t)$, leniency is applied to a Q-value update as follows:

$$Q(s_t, a_t) = \begin{cases} Q(s_t, a_t) + \alpha \delta, & \delta > 0 \text{ or } x > l(s_t, a_t). \\ Q(s_t, a_t), & \delta \leq 0 \text{ and } x \leq l(s_t, a_t). \end{cases} \quad (5)$$

A random variable $x \in [0, 1]$ that draws samples from a uniform distribution is used to ensure that an update on a negative δ is executed with a probability $1 - l(s_t, a_t)$. In repeated games the temperature values for state transitions close to areas where the agents are spawned can decay rapidly, due to initial state-action pairs being visited more often than the later ones. It is crucial however for the success of the lenient learners that the temperatures for these states-action pairs remains sufficiently high for the rewards to propagate back from the later stages, and to prevent the agents from converging upon a sub-optimal policy. One solution to this problem is to fold the average temperatures for the n actions available to the agent in s_{t+1} into the temperature that is being decayed for (s_t, a_t) , where the average temperature value is calculated as follows: $\bar{T} \leftarrow 1/n \sum_{i=1}^n T(s_{t+1}, a_i)$ [15]. The extent to which \bar{T} is folded in is determined by a constant ν using Equation 6:

$$T_{t+1}(s_t, a_t) = \beta \times \begin{cases} T_t(s_t, a_t), & \text{if } s_{t+1} \text{ is terminal.} \\ (1 - \nu)T_t(s_t, a_t) + \nu\bar{T}_t(s_{t+1}), & \text{otherwise.} \end{cases} \quad (6)$$

3 Algorithmic Contributions

In this section we describe our two new algorithms, i.e., the Lenient Deep Q-Network and the Scheduled Hysteretic Deep Q-Network:

Lenient Deep Q-Network (LDQN). Our approach towards applying leniency to MA-DRL is to store the temperature value associated with each state-transition in the experience replay memory: $(s_{t-1}, a_{t-1}, r_t, s_t, \tau_t)$, where $\tau_t \leftarrow T(s_{t-1}, a_{t-1})$. Therefore τ_t is used to calculate the amount of leniency that should be used when a transition is sampled, with $l(\tau_t) \leftarrow 1 - e^{-K*\tau_t}$. To reduce memory consumption the temperatures are stored by generating a hash key for states that have been visited, and then mapping this key with the action selected to the temperature value.

As in the previous section the aim is to minimize the loss function outlined in equation 3, with the modification that for each sample j chosen from the replay memory, δ_j is set to 0 if the conditions outlined in Equation 5 are not met. We find that maintaining a slow-decaying global temperature ν helps stabilize the learning process. Without the global temperature the disparity between the low temperatures in well explored areas and the high temperatures in relatively unexplored areas has a destabilizing effect during the later stages of the learning process. The global temperature ν is stored whenever $\nu_t < T(s_{t-1}, a_{t-1})$.

Temperature Decay Schedule (TDS). Despite implementing lenient agents using average temperature folding (ATF) we find that temperatures decay rapidly for state-action pairs belonging to challenging sub-tasks in the environment. As a result we developed an alternative approach using a pre-computed temperature decay schedule $\beta_{0, \dots, MaxSteps}$, where b_0 is given an initial decay value, and for each $b_n \in \beta$ where $n > 0$, $b_n \leftarrow e^{-2*b_0^{d^n}}$, where d is a decay rate. Upon reaching a terminal state the temperature decay schedule is applied as follows to ensure that temperature values belonging to state-action pairs encountered during the early phase of an episode are decayed at a slower rate than those close to the terminal state-transition:

Algorithm 1 Application of temperature decay schedule (TDS)

```
1:  $n \leftarrow 0$  and  $steps \leftarrow$  steps taken during the episode
2: for  $i = steps$  to 0 do
3:    $T_{t+1}(s_i, a_i) \leftarrow \beta_n T_t(s_i, a_i)$ 
4:    $n \leftarrow n + 1$ 
5: end for
```

Scheduled Hysteretic-DQN (SHDQN). Optimistic learners often converge towards sub-optimal joint policies in environments that yield stochastic rewards [15]. However, drawing parallels to lenient learning, where it is desirable to decay state-action pairs encountered at the beginning of an episode at a slower rate compared to those close to a terminal state, we consider that the same principle can be applied to hysteretic Q-learning. Subsequently we implemented a Scheduled Hysteretic-DQN with a pre-computed schedule $\beta_{0 \dots MaxSteps}$ where $b \in \beta$ is set to a value approaching α , and for each $b_n \in \beta$ where $n > 0$, $b_n \leftarrow d \times b_{n-1}$ using a decay coefficient d where $d \leftarrow (0, 1]$. The state transitions encountered throughout each episode are initially stored within a queue data-structure. Upon reaching a terminal state the n state-transitions are transferred to the experience replay memory as $(s_t, s_{t+1}, r_{t+1}, a_t, b_{n-t})$ for $t \leftarrow 0$ to n . Our hypothesis is that storing β values that approach α for state-transitions leading to the terminal state will help agents converge towards the optimal joint policy in environments that yield a stochastic reward.

4 Experimental Evaluation

Coordinated Multi-Agent Object Transportation Problem (CMOTP). We subjected our agents to a range of CMOTPs inspired by the scenario discussed in Buşoniu et al. [3]. Two agents are tasked with delivering one item of goods to a drop-zone within a grid-world. The agents must first exit a room one by one before locating and picking up the goods by standing in the grid cells on the left and right hand side. The task is fully cooperative, meaning the goods can only be transported upon both agents grasping the item and choosing to move in the same direction. Both agents receive a positive reward after placing the goods inside the drop-zone. The actions available to each agent are to either stay in place or move left, right, up or down. We subjected agents to three variations of the CMOTP, depicted in figure 1, where each A represents one of the agents, G the goods, and $D-ZONE / DZ$ marks the drop-zone(s). The layout in sub-figure 1a is a larger version of the original layout [3], while the layouts in sub-figures 1b and 1c introduce narrow-passages between the goods and the drop-zone, testing the agents ability to learn to cooperate in order to overcome challenging areas within the environment. The layout in sub-figure 1c additionally tests agents' response to stochastic rewards. Drop-zone 1 (DZ1) yields a reward of 0.8, whereas drop-zone 2 (DZ2) returns a reward of 1 on 60% of occasions and only 0.4 on the other 40%. DZ1 therefore returns a higher reward on average, 0.8 compared to the 0.76 returned by DZ2.

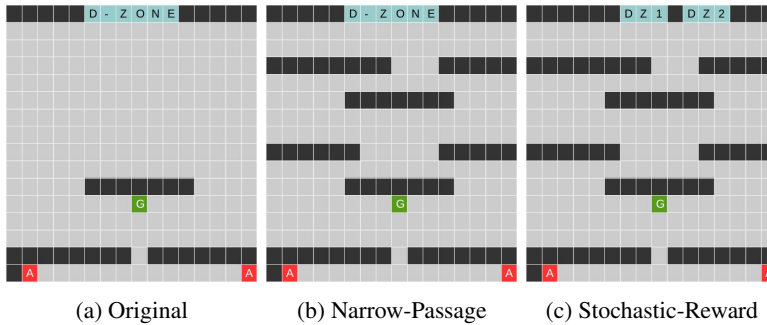


Figure 1: Coordinated Multi-Agent Object Transportation Problem (CMOTP) Layouts

Experiments We conducted 10 runs of 5000 episodes for each algorithm, with a limit of 10'000 time-steps per episode. Our DQN consisting of 2 convolutional layers with 32 and 64 kernels respectively, a fully connected layer with 1024 neurons and an output neuron for each action. Work

conducted by Gupta et al. [13] inspired us to represent the state-space as image like tensors with 4 channels. As a result each agent is fed a $16 \times 16 \times 4$ tensor with channel 1 being dedicated to obstacles, channel 2 to the goods, channel 3 to the teammate and channel 4 providing the agent’s location. The tensors are sparse. For example all cells within channel 2 are set to 0 with the exception of the one cell that contains the goods, which is set to 1. *Adam* [24] was used to optimize the networks with $\alpha = 0.0001$ and $\gamma = 0.95$. Target network synchronization takes place every 5000 steps for double q-learning. An epsilon greedy exploration strategy is utilized, with $\epsilon \leftarrow 1$ initially, before being decayed after each episode with a factor of 0.999. The minimum value for ϵ is 0.05. The replay memory size is set to 250’000 state transitions. The values of the hyperparameters were chosen based upon agents delivering strong performances while solving a maze task and being subjected to CMOTP test runs. For the lenient agents we tried a number of temperature modification coefficients K before settling upon the value of 2. For the initial temperature values for state-actions pairs we find that 1 works well, delivering the right amount of leniency to ensure that some negative updates do take place. During initial trials it became apparent that agents with too much leniency fail to converge due to the unchecked growth of the Q-values. Finally we pre-compute the temperature decay schedule using $d \leftarrow 0.9$ and $b_0 \leftarrow 0.005$. SHDQNs are initialized with $\beta_0 = 0.9$ and a decay coefficient of $d = 0.99$ being applied while $\beta_n > 0.4$.

5 Results

Original CMOTP. Standard DQN and Double-DQN architectures struggled to master the CMOTP, with only 4 and 8 runs respectively converging towards an efficient joint policy. The lenient and hysteretic architectures with $\beta < 0.8$ fared significantly better, converging towards joint policies that were only a few steps shy of the optimal 33 steps required to solve the task. After analyzing the results from the initial 10 runs we noticed that while standard Q-learning implementations outperformed their double Q-learning counterparts throughout the initial 1000 episodes, as evident in Figure 2, double Q-learners tended to converge towards superior joint policies. To test the significance of these results we decided to conduct a further 20 runs for each of the lenient and hysteretic settings. We subsequently performed a Kolmogorov-Smirnov test with a null hypothesis that there is no significant difference between the standard and double Q-learning implementation for each network configuration.

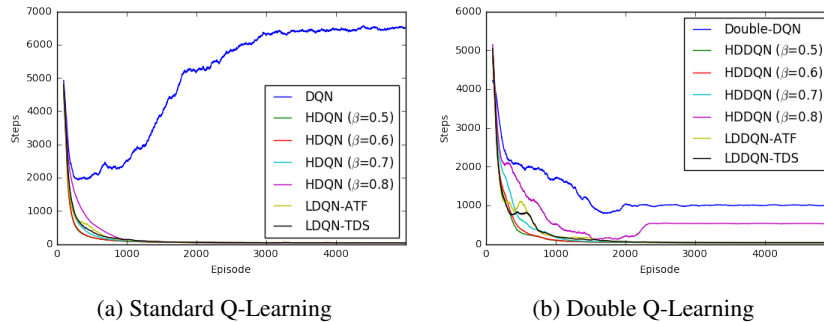


Figure 2: Original CMOTP - Average steps per episode

We evaluated the agents based upon the average steps required to complete the task and the percentage of coordinated steps over the final 100 episodes. As evident by the p-values listed in Table 1 the null hypothesis was rejected for each network setting, with double Q-learners having a significantly higher coordinated steps percentage and lower average steps per episode. The findings inspired us to implement hybrid agents that use vanilla Q-learning over the first n episodes before converting to double Q-learning. The performance of the hybrid agents is discussed in the following sections.

	Coordinated steps percentage			Average steps per episode		
	Q-Learning	Double Q-Learning	P-Values	Q-Learning	Double	P-Values
Hysteretic $\beta = 0.5$	0.91253	0.92795	3.3112e-06	37.951	36.47214	0.01054
Hysteretic $\beta = 0.6$	0.91662	0.92566	0.00061	38.062	36.13866	0.00061
Hysteretic $\beta = 0.7$	0.91595	0.92307	0.01131	39.573	36.87933	0.02585
Lenient ATF	0.89842	0.92516	8.3835e-12	40.404	36.98266	2.6198e-07
Lenient TDS	0.91540	0.92188	0.01131	37.615	36.79	0.01131

Table 1: Average steps per episode and coordinated steps percentage over the final 100 episodes for experiments conducted in the original CMOTP

Narrow-Passage CMOTP. Lenient agents implemented with a TDS performed consistently across Q-learning, double Q-learning and hybrid runs conducted within the Narrow-Passage CMOTP, with a coordinated steps percentage of around 92% and averaging just over 45 steps per episode. The other algorithms lacked this consistency as evident by the results averages for HDQN ($\beta = 0.5$) and HDDQN ($\beta = 0.6$) in Table 2. Meanwhile the hybrid agents, who switched from standard to double Q-learning after 1500 episodes, converged towards efficient cooperative policies in each setting. However, the fact that the majority of vanilla Q-learning trials required less steps per run on average hints at the existence of a superior point for the Hybrid agents to transition from vanilla to double Q-learners.

	QL	Double QL	Hybrid	QL	Double QL	Hybrid	QL	Double QL	Hybrid
	Coordinated steps percentage			Average steps per episode			Average steps per run		
Hyst. $\beta = 0.5$	0.839	0.925	0.924	1053.669	44.84	45.32	1888279.8	1602246.9	1326856.9
Hyst. $\beta = 0.6$	0.884	0.888	0.921	86.805	1030.56	47.75	1841199.6	6163874.1	2101365.8
Lenient-ATF	0.905	0.924	0.923	49.811	46.862	46.18	1886812.0	2516973.9	1962465.4
Lenient-TDS	0.918	0.923	0.923	45.513	45.164	45.72	1582855.4	2051219.7	1603617.3

Table 2: Narrow-Passage CMOTP Results

Table 2 shows that lenient agents implemented with a TDS outperformed the ATF agents for vanilla, double Q-learning and hybrid configurations. A likely contributing factor to the ATF agents converging on less optimal joint policies is the premature decay of temperature values for state-actions pairs within the first two compartments of the narrow-passage CMOTP. Figure 3 contains two heat-maps illustrating the average temperature values for each cell within the grid for agents holding the goods. The values are obtained by averaging the temperatures stored for each action associated with the individual states visited. The heat-maps were generated after only 700 episodes, at which point temperatures have decayed significantly within the first and second compartments for agents implemented with ATF. Meanwhile the temperature values decayed using a TDS are lower towards the terminal transition, ensuring that the agents apply sufficient leniency while the rewards are slowly propagated back from later steps.

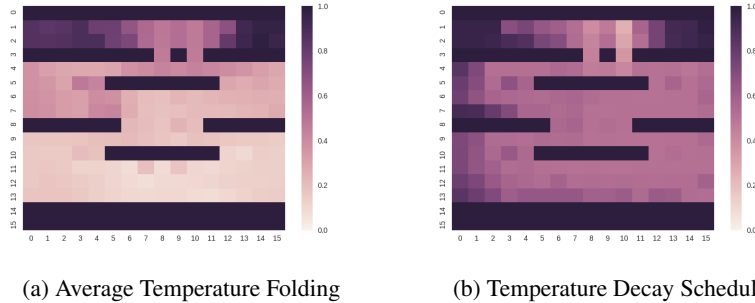


Figure 3: Average temperature comparison after 700 episodes in the Narrow-Passage CMOTP

Stochastic Reward CMOTP. Facing the stochastic reward CMOTP the LDQN-TDS and its hybrid version significantly outperformed hysteretic DQNs with a fixed β value. A large proportion of the runs conducted for Lenient-ATF and Hysteretic agents with $0.5 < \beta$ failed to converge upon a cooperative policy. Meanwhile Hysteretic Q-learners with $\beta = 0.5$ only converged upon a policy that consistently delivered the goods to the sub-optimal drop-zone 2. Agents implemented with the LDQN-TDS meanwhile converged upon a joint policy that resulted in the goods being delivered to the optimal drop-zone, as evident from the average reward of 0.8 obtained over the final 100 episodes, listed in Table 3. The only other algorithm that delivered competitive results was the SHDQN. However, the SHDQN runs lacked stability, as evident from the running average reward and coordinated steps percentages plotted in figure 4. Furthermore the first run failed to converge for the Hybrid version of the SHDQN. Upon restarting the experiment the setting delivered 10 straight successful runs. There may therefore exist a more optimal β schedule that could increase stability.

Algorithm	Average steps	Coordinated steps percentage	Average reward
HDQN $\beta = 0.5$	108.07899	0.86251	0.7504
Hybrid HDQN $\beta = 0.5$	59.523	0.87505	0.7546
SHDQN	50.967	0.91032	0.7938
Hybrid SHDQN	70.367	0.91006	0.795
LDQN-TDS	48.49199	0.91971	0.8
Hybrid LDQN-TDS	53.561	0.90764	0.7966

Table 3: Stochastic Reward CMOTP: Results from final 100 episodes

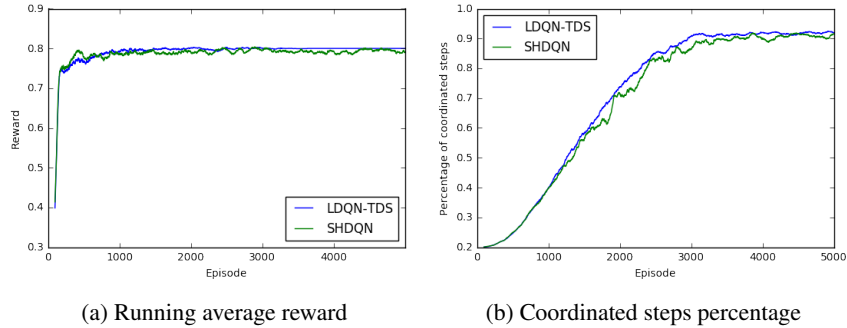


Figure 4: Comparison of LDQN-TDS and SHDQN in the stochastic reward CMOTP

6 Discussion & Conclusion

We have introduced a lenient deep reinforcement learning algorithm and have shown that it can help two agents master an episodic fully cooperative transportation task that yields stochastic rewards. Furthermore, we have demonstrated the advantage of using a temperature decay schedule to prevent the fast decay of temperature values for state-action pairs belonging to state transitions encountered within the early phases of an episode. Following the successful application of our temperature decay schedule to leniency we applied a similar concept to hysteretic Q-learning, achieving encouraging results. Applying less optimism to updates for state-transitions close to a terminal state enabled our SHDQN agents to significantly outperformed standard HDQN agents in a domain that yielded a stochastic reward at the end of each episode.

Finally we compared the performance of hysteretic and lenient agents implemented with vanilla and double Q-learning algorithms. We found support that vanilla Q-learners converge at a faster rate, while double Q-learners converge towards a superior joint policy. These results do not come as a surprise, given that double Q-learning is intended to reduce overoptimistic value estimates, thereby initially hampering the learning process. Meanwhile the results support that using double Q-learning at a later phase can help the agents converge towards a superior joint policy. As a result we advocate a hybrid approach, where the cooperative agents initially use a vanilla Q-learning approach before

converting to double Q-learners at a later stage. In this current work the transition to double Q-learning took place at a defined time. However, in future research we plan to investigate whether a more informed approach is possible for choosing the conversion point.

We see the application of leniency towards MA-DRL as a two step task. The first step involves enabling agents within a cooperative multi-agent setting to be implemented with leniency whilst sampling transitions from an experience replay memory. The second task involves mapping temperature values required by leniency to a continuous state-action pair. In this work we have addressed how the first task can be solved through storing the temperature value for a state-action pair at time t with the state-transition. Our future research will therefore focus on the second step, exploring methods to allow our lenient deep reinforcement learners to cope with continuous state and action spaces within a parallel MA-DRL context. One potential approach that we plan to explore is to train the agent to decide how much leniency to apply to a given state-transition during an update, with the amount of leniency returned being implemented as an auxiliary control task [25].

Acknowledgments

We thank the HAL Allergy Group for partially funding the PhD of Gregory Palmer and gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU that enabled this research.

References

- [1] J. Foerster, N. Nardelli, G. Farquhar, P. Torr, P. Kohli, S. Whiteson, et al. Stabilising experience replay for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1702.08887*, 2017.
- [2] L. Panait, K. Sullivan, and S. Luke. Lenient learners in cooperative multiagent systems. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 801–803. ACM, 2006.
- [3] L. Buşoniu, R. Babuška, and B. De Schutter. Multi-agent reinforcement learning: An overview. In *Innovations in multi-agent systems and applications-I*, pages 183–221. Springer, 2010.
- [4] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian. Deep decentralized multi-task multi-agent rl under partial observability. *arXiv preprint arXiv:1703.06182*, 2017.
- [5] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
- [6] G. Lample and D. S. Chaplot. Playing fps games with deep reinforcement learning. In *AAAI*, pages 2140–2146, 2017.
- [7] H. Van Hasselt, A. Guez, and D. Silver. Deep reinforcement learning with double q-learning. In *AAAI*, pages 2094–2100, 2016.
- [8] T. Schaul, J. Quan, I. Antonoglou, and D. Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.
- [9] T. de Bruin, J. Kober, K. Tuyls, and R. Babuška. The importance of experience replay database composition in deep reinforcement learning. In *Deep Reinforcement Learning Workshop, NIPS*, 2015.
- [10] S. Gu, E. Holly, T. Lillicrap, and S. Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. *arXiv preprint arXiv:1610.00633*, 2016.
- [11] L. Busoniu, R. Babuska, and B. De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, And Cybernetics-Part C: Applications and Reviews*, 38 (2), 2008, 2008.
- [12] K. Tuyls and G. Weiss. Multiagent learning: Basics, challenges, and prospects. *AI Magazine*, 33(3):41–52, 2012.
- [13] J. K. Gupta, M. Egorov, and M. Kochenderfer. Cooperative multi-agent control using deep reinforcement learning. In *Proceedings of the Adaptive and Learning Agents workshop (at AAMAS 2017)*, 2017.
- [14] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, and T. Graepel. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.
- [15] E. Wei and S. Luke. Lenient learning in independent-learner stochastic cooperative games. *Journal of Machine Learning Research*, 17(84):1–42, 2016. URL <http://jmlr.org/papers/v17/15-417.html>.
- [16] L. Panait, K. Tuyls, and S. Luke. Theoretical advantages of lenient learners: An evolutionary game theoretic perspective. *Journal of Machine Learning Research*, 9(Mar):423–457, 2008.

- [17] A. Barto and R. Sutton. *Reinforcement learning: An introduction*. MIT press, 1998.
- [18] H. V. Hasselt. Double q-learning. In *Advances in Neural Information Processing Systems*, pages 2613–2621, 2010.
- [19] L. Matignon, G. J. Laurent, and N. Le Fort-Piat. Hysteretic q-learning: an algorithm for decentralized reinforcement learning in cooperative multi-agent teams. In *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, pages 64–69. IEEE, 2007.
- [20] M. A. Potter and K. A. De Jong. A cooperative coevolutionary approach to function optimization. In *International Conference on Parallel Problem Solving from Nature*, pages 249–257. Springer, 1994.
- [21] D. Bloembergen, D. Hennes, M. Kaisers, and K. Tuyls. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53:659–697, 2015. ISSN 10769757.
- [22] R. Sarin and T. Börgers. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77(1):1–14, 1997.
- [23] K. Tuyls and A. Nowé. Evolutionary game theory and multi-agent reinforcement learning. *The Knowledge Engineering Review*, 20(01):63–90, 2005.
- [24] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.
- [25] M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*, 2016.